

**Development Phase – Use Case Analysis.**

**M.Sc. Data Science.**

**DLMDSPDSUC01 – Project : Data Science Use Case**

**13th April 2023.**

**Caroline Karimi Gitari.**

**92125360.**

**Tutor: Prof. Sahar Qaadan.**

# **Introduction to the Machine Learning Canvas**

The machine learning canvas can be described as consisting of 4 main blocks:

* Value proposition – Includes : What are we trying to do, why is it important and who is going to be impacted by it.
* Learn - It contains the following boxes: Data sources, collecting data, Features, and Building models.
* Predict – Based on the machine learning models to be used to learn from data.
* Evaluate – Involves evaluating how well the model is performing.

Figure 1 below shows the Machine learning canvas blocks and what they include.



Figure 1: Machine Learning Canvas Blocks

# **Bora Credit Use Case Analysis**

## **Value proposition**

Bora Credit is a digital lending startup that aims to provide loans to online merchants to help them purchase enough stock. To achieve this goal, the company currently uses an expert-based credit model to assess creditworthiness. However, using historical data from the first loan cycle, the company aims to build a model that can predict whether to lend to the customer or not and how much money to lend. The use of machine learning in creditworthiness prediction has the potential to significantly improve the lending process for Bora Credit and its customers.

The value proposition for the prediction model includes:

* **Accurate and Consistent Credit Decisions** -The use of a machine learning algorithms to predict creditworthiness will enable Bora Credit to make more accurate and consistent credit decisions. The model will be trained on historical data from the previous loan cycle to identify patterns and factors that indicate creditworthiness or risk. By using a machine learning model, Bora Credit can make credit decisions that are based on data-driven insights rather than subjective judgments which are more prone to errors. This will reduce the risk of lending to customers who are unlikely to repay their loans and increase the likelihood of lending to those who will.
* **Reduced Manual Effort and Costs** - The current expert-based credit model requires significant manual effort to evaluate creditworthiness. This process is time-consuming and costly and continues to get more challenging with each increasing number of loan applications. The machine learning model will automate the credit evaluation process, reducing the manual effort required and lowering costs. This will enable Bora Credit to evaluate more loan applications quicker and with less manpower, reducing the cost of lending for the company.
* **Improved Customer Experience –** With the increasing number of daily credit applications, there have been several customer complaints about the loan approval times and delays. The new model will improve the customer experience for Bora Credit's customers in the following ways :
  + Faster and more efficient decisions - By automating the credit evaluation process, customers can receive loan decisions faster and more efficiently. Our target customers are small business owners who rely on quick access to capital to purchase stock and keep their business running smoothly, hence the speed of loan processing is extremely important to them.
  + Accurate credit decisions – Machine learning will help Bora Credit reduce the likelihood of customers defaulting on their loans, which can negatively impact their credit score and future borrowing ability.
* **Increased Lending Volume and Revenue -**The machine learning model will enable Bora Credit to increase its lending volume and revenue by:
  + More accurate credit decisions - Lending to customers who are likely to repay their loans.
  + Expansion of the customer base - Some customers who may have been wrongly rejected previously will now be approved.

## **Learn**

This section focuses on the learning aspect of the Machine Learning Canvas. The objective is to understand the data sources, and the data collection process, identify the features and build models for creditworthiness prediction.

### **Data Sources**

Here, we aim to answer the question, "Which raw data sources can we use?"

Bora Credit has already collected data from its first loan cycle. The data collected includes:

* Categorical data:
  + Gender.
  + Business registration type.
  + Saving frequency.
  + Quality checks.
* Numerical Data :
  + Monthly Sales.
  + Monthly turnover.
  + Businesses Owned.
  + Total employees.

This data will serve as the primary source for building the prediction model.

### **Collecting Data**

First loan cycle data is readily available and will be exported to an excel sheet and restructured so that each row represents a unique customer. In addition to self-reported merchant data, Bora credit has also included a CRB integration that allows the extraction of credit scores and the history of previous loans taken by the merchants.

To improve the model's accuracy, some additional data from various sources include:

* Additional variables generated from existing historical data.

### **Features**

Features are input representations extracted from raw data sources. Any variables with a significant number of missing values will be dropped altogether.

The features that we will use include:

* Gender.
* Business registration type.
* Saving frequency.
* Quality checks.
* Businesses’ competitive advantage.
* Business premises.
* CRB band.
* Education level.
* Product categories sold.
* Total delayed supplier payments.
* Number of product categories sold.
* Monthly Sales.
* Monthly turnover.
* Number of Businesses Owned.
* Total employees.
* Number of current loans.
* Number of financial dependants.
* Number of active loans owed.
* Number of addresses over the last 5 years.
* Total CRB enquiries last 12 months.
* Total phone numbers over the last 5 years.

In addition to the raw variables from customer historical loan data, additional variables will be generated. For instance, using the total number of loans borrowed and the total number of loans defaulted, a new variable, the share of loan defaults can be calculated.

2 outcome variables will be used for prediction:

* Loan status – business classification of the loan.
* Amount Paid – Total amount repaid by the customer against the loan.

### **Building Models**

The applicable machine learning methodology includes :

* Regression - Amount paid.
* Classification – Loan status.

Figure 2 shows the models to be used for each outcome variable.

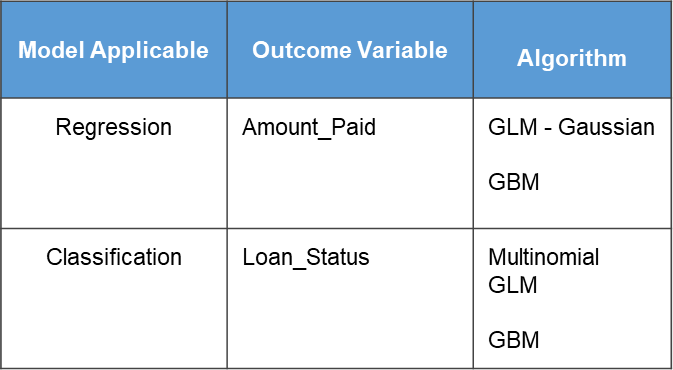


Figure 2:ML Models Bora Credit

GLM-Gaussian is a linear regression model that assumes a Gaussian (normal) distribution for the response variable, while GBM is a tree-based ensemble model that uses gradient boosting to improve performance.

The advantages of our model selections include:

* GLM-Gaussian for predicting amount paid.
  + It is a simple and interpretable model.
  + It assumes a normal distribution for the response variable, which is appropriate when the data follows a normal distribution.
* GBM for predicting amount paid.
  + It is a more complex model that can handle nonlinear relationships and interactions between variables.
  + It may perform better than GLM-Gaussian if the data has nonlinear relationships.
* Multinomial GLM for predicting loan status.
  + It is a generalized linear model that allows for multiple categories in the response variable. It assumes that the response variable follows a multinomial distribution.
* GBM for predicting loan status.
  + It is a tree-based ensemble model that can handle multiple categories.

By running two models for each outcome variable, we can compare their performance and choose the model that performs better. It is not always clear which model will perform better for a given dataset, so it is important to try different models and compare their accuracy. Additionally, by running two different models, we can check if the results are consistent across different modeling techniques.

### **Conclusion**

In conclusion, the learn phase of the machine learning canvas has provided us with a clear understanding of the data sources, data collection process, features required, and the machine learning algorithms required for building the prediction model. In the next phase, we will focus on the predict phase of the machine learning canvas, where we will train the model, test its accuracy, and deploy it in a production environment.

## **Predict**

The left-hand part of the machine learning canvas (PREDICT) contains the following boxes: ML Task, Decisions, Making Predictions, and Offline Evaluation.

### **ML Task**

Machine learning tasks can be thought of as specific questions which the model aims to answer. The questions must be specific and for Bora credit can be considered as :

* What will be the loan status for this new client in the future?
* How much will this new client repay against their loan?

The data sources in the learn section provide machine learning tasks that are mainly characterized by the Input, output to predict, and type of problem.

* **Input –** Categorical and numerical data on the merchants from their first loan cycle.
* **Output** - Amount paid and loan status. The amount paid is a continuous variable that represents the total amount paid back by the borrower, while loan status is a categorical variable that indicates whether the loan was paid in full, partially paid, or defaulted. It is important to check the output variable distribution and range to handle any imbalances and misrepresentations.
* **Type of Problem**- The problem is a supervised learning problem, as we have labeled data and aim to build a model that can predict the creditworthiness of future loan applicants.

### **Decisions**

Bora credit will be making 1 prediction per customer for around 1000 customers monthly. The predictions will be used to make decisions about whether to lend money to a customer and if so, how much to lend. The proposed value to the end user is that Bora Credit can make better lending decisions and reduce the risk of defaults. The final lending decision will be made by comparing the predicted creditworthiness of the customer to Bora Credit's risk tolerance level. If the predicted creditworthiness meets the risk tolerance level, Bora Credit will approve the loan application.

### **Making Predictions**

Some technical constraints that will need to be considered while making predictions to support decisions include:

* Prediction volume – Bora credit has currently projected around 1,000 new customers monthly.
* Time constraint – The model needs time to extract features and make predictions. Featurization will take a bit longer as it involves accessing various databases and performing feature extractions. On the other hand, customers require real-time predictions to enable them to purchase stock. The time taken for each prediction will be a tradeoff between the time required for the application and what is acceptable for the end user.
* System robustness to loss of user internet connectivity – if we want to make predictions directly on the user device, we might need a system that is robust to loss of internet connectivity. If devices are not powerful enough to run predictions on time, then this constraint will point toward cloud usage.

### **Offline Evaluation**

Before deploying the system, we need to evaluate the performance of the model. In addition to model accuracy, we will need to pre-evaluate the future impact of decisions to gain confidence that the model is ready for deployment.

**Test Cases**

We will split the dataset into training and testing sets, where the training set will be used to train the model, and the testing set will be used to evaluate the model's performance. For the model to be trustworthy, the test cases used should be representative of what it will encounter once deployed. Similarly, the training data should be representative of what the model will be tested on. Contrary to the commonly used method where the test set and train set are split randomly which may lead to overly optimistic evaluation results, we will apply domain and data knowledge to decide on the most efficient segregation.

**Test Size**

Since we currently do not have a lot of data, the old heuristic of a 70-30 train test is sufficient. This may change in future, as the test size can significantly affect the prediction/decision time.

**Metrics**

For performance evaluation, it is important to select a domain-specific metric that helps quantify the effect of the model. Instead of only using abstract model performance metrics (e.g., the regression models Mean Absolute Error), we will use these metrics to compute the company’s gain (or cost reduction) by using the model compared to not using it.

The performance metrics we will use include:

* Mean absolute error (MAE) for the regression models – The model with the lower MAE between GLM – Gaussian and GBM will be selected as the ideal model.
* Precision, recall, and overall accuracy of the classification models – The model that performs better will be selected as ideal.
* Bora Credit Gain from using the new model.
* Mean amount of bad loans with the current expert model vs the data-based prediction model

Figure 3 shows the sample structure and some metrics we will use to assess the impact of the application of the new model to Bora credit current loans.

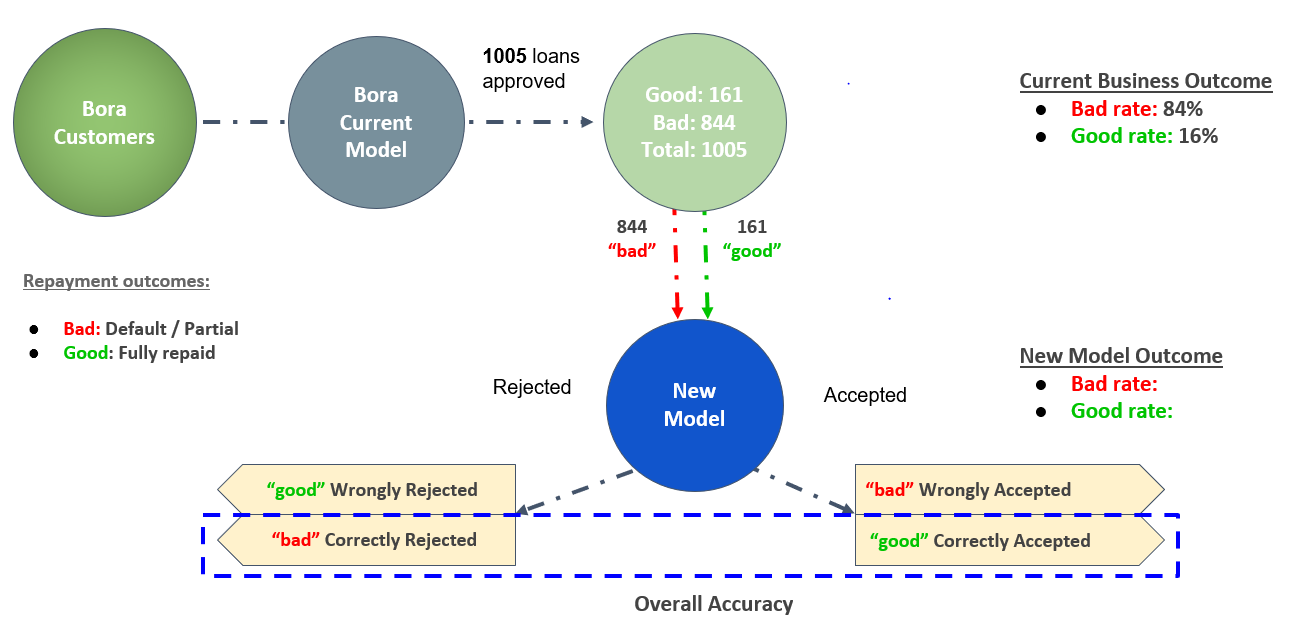


Figure 3: New model assessment

Once we have evaluated the model's performance, we can deploy it to make predictions on new data.

## **Evaluate**

Model development and deployment is not a one-time activity, but rather an ongoing one that will require constant monitoring and evaluation. To ensure that the model remains effective and accurate over time, it is important to continuously evaluate its performance and adjust as necessary. This can be accomplished through a variety of means, such as regular data analysis, model validation, and performance metrics tracking.

### **Life Evaluation and Monitoring**

To ensure that the prediction model remains effective and relevant over time, Bora Credit will regularly monitor its performance using :

* Regular data analysis
* Performance metrics tracking
* Key performance indicators – these will be designed to measure the effectiveness of the model in accurately predicting creditworthiness, as well as its impact on the overall lending process. Some examples of KPIs include:
  + Number of loans approved or denied based on the model's predictions.
  + Percentage of loans that are paid back in full.
  + Overall profitability of the lending program.

In addition to monitoring the performance of the model, Bora Credit will also consider how the model is being used in practice which will include:

* Gathering feedback from borrowers to identify areas where the model could be improved or refined.
* Regular communication and collaboration between the data science team and the lending team to ensure that the model is aligned with the company's overall business objectives.

By regularly evaluating the performance of the model and monitoring its impact on the lending process, Bora Credit will ensure that it is making data-driven lending decisions that are both effective and sustainable over time.

# **Filled in ML Canvas**

